# NHẬP MÔN KHOA HỌC DỮ LIỆU

**TS. NGUYỄN HUYỀN CHÂU**
**AI LAB - TLU**

# L01a: Tiền xử lý và EDA

Tiền xử lý dữ liệu (tiếp)

Phân tích cơ bản

Thống kê mô tả

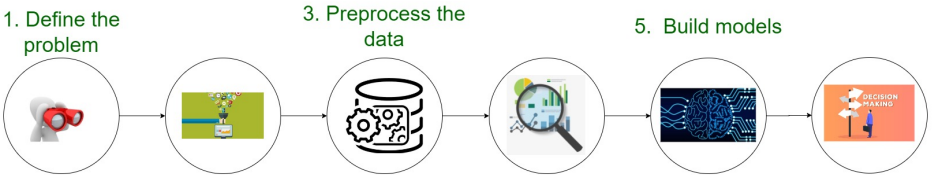
# Outline

**Tiền xử lý dữ liệu (tiếp)**

Phân tích cơ bản

Thống kê mô tả

# Tiền xử lý dữ liệu




Đọc
Làm sạch

# Tiền xử lý dữ liệu

Chuẩn bị dữ liệu để thuận tiện cho các phân tích thống kê ở bước sau

- **Đọc dữ liệu:** chuyển dữ liệu vào chương trình, trong đó có cấu trúc hoá dữ liệu (kiến trúc, định dạng, logic). Việc đọc sẽ trừu tượng hoá dữ liệu thành các đơn vị dễ hiểu thuận tiện xử lý xử lý về sau.
- **Làm sạch dữ liệu**: xử lý các vấn đề tiềm ẩn trong dữ liệu, đảm bảo dữ liệu sẵn sàng cho các bước khai phá sau này

# Làm sạch dữ liệu

Xử lý các vấn đề tiềm ẩn để dữ liệu sẵn sàng phân tích, xử lý về sau

- **“Gabage in, gabage out” (Vào rác, ra rác)** là quy luật cảnh báo trong khoa học dữ liệu. Dữ liệu mang vấn đề tiềm ẩn có thể làm hỏng toàn bộ kết quả về sau. Cần đảm bảo rác không lọt vào ngay từ đầu

Các vấn đề tiềm ẩn: thiếu hụt, sai sót, xung đột, bất thường

- Giá trị thiếu
- Lỗi vs. Giả tượng
- Sự không tương thích
- Giá trị bất thường

Lưu ý: Backup bản chính và làm sạch chỉ trên bản sao

# Giá trị thiếu

- Các tập dữ liệu có thể thiếu giá trị vì nhiều lý do
  - Dữ liệu vốn không có. VD: năm mất của người đang sống
  - Dữ liệu có mà ít. VD: tỷ lệ người mắc bệnh xương thuỷ tinh
  - Do người khảo sát quên hay cung cấp dữ liệu sai. Vd: năm sinh 1890
- Đôi khi xử lý các vấn đề khác (trùng lặp, lỗi logic) bằng cách bỏ giá trị tại cột gây vấn đề, cần bổ sung giá trị khác và quy về xử lý giá trị thiếu
- Cách xử lý giá trị thiếu đòi hỏi nhiều cân nhắc:
  - **Tìm nguồn bổ sung dữ liệu**
  - **Tự bổ sung giá trị**: ưu tiên cách này nếu khả thi
  - **Bỏ dòng:** yêu cầu dữ liệu còn lại vẫn đủ về lượng, chất (đủ tính đa dạng), thường chỉ khi dữ liệu thiếu không mang tính hệ thống
  - **Bỏ cột:** chỉ khi có quá nhiều giá trị thiếu cho cột đó

# Bổ sung giá trị thiếu

Không nên bổ sung bằng giá trị rỗng hay giá trị vô lý (vd -1 cho tuổi) bởi các mô hình phân tích có thể vô tình tính toán cả các giá trị này. Hãy ước lượng ra giá trị bổ sung phù hợp từ các giá trị đã có:

- **Ước lượng theo kinh nghiệm:** năm mất = năm sinh + 80
- **Ước lượng theo trung bình:**
  - Không làm thay đổi giá trị trung bình tổng,
  - không ảnh hưởng gì đến kết quả chung,
  - không dùng được nếu các giá trị thiếu mang tính hệ thống. Vd: năm mất

# Giá trị thiếu

- **Ước lượng theo ngẫu nhiên**: lấy trong các giá trị đã có ở cột. Đánh giá chất lượng mô hình ngẫu nhiên bằng cách chạy nhiều lần
- **Ước lượng theo láng giềng gần nhất**: lấy giá trị từ dòng giống dòng bị thiếu nhất, cho dữ liệu đa dạng hơn lấy trung bình.
- **Ước lượng theo ngoại suy**: Xây dựng mô hình để dự đoán giá trị cột bị thiếu từ giá trị những cột khác. Vd: mô hình hồi quy tuyến tính
  - Chỉ dùng khi mỗi hàng thiếu 1 giá trị, lạm dụng sẽ tạo ra các bản ghi bất thường, kéo phân tích ngả về quy luật của các hàng thiếu giá trị

# Ví dụ 1

| Ngày | Nhiệt độ trung bình | Nhiệt độ cao nhất | Lượng mưa |
|---|---|---|---|
| 01/01/2021 | 30 | 32 | 160 |
| 02/01/2021 | 34 | 31 | |
| 03/01/2021 | | 26 | 145 |
| 04/01/2021 | Hai sáu | 28 | 122 |
| 04/01/2021 | 26 | 21 | 111 |

**Vấn đề thiếu dữ liệu:**

- Tìm nguồn dữ liệu thay thế trong bảng hoặc dữ liệu từ một nguồn khác
- Nếu không thể, thực hiện bổ sung dữ liệu dùng các kỹ thuật bổ sung

**Vấn đề dữ liệu sai logic:**

- Loại bỏ dữ liệu gặp vấn đề
- Thay thế bằng giá trị phù hợp để hai cột không còn gặp vấn đề (câu hỏi: nên chọn cột nào là sai để xoá và sửa?)

**Vấn đề về dữ liệu bị trùng lặp:**

- Thực hiện kiểm tra và thay thế giá trị thông qua 1 nguồn khác (Có thể dự đoán ngày 04/01/2021 bị lặp là ngày 05/01/2021 từ quy luật về ngày tháng trong bảng)
- Nếu không thể, loại bỏ dữ liệu bị lặp và chỉ giữ lại 1

**Vấn đề format:**

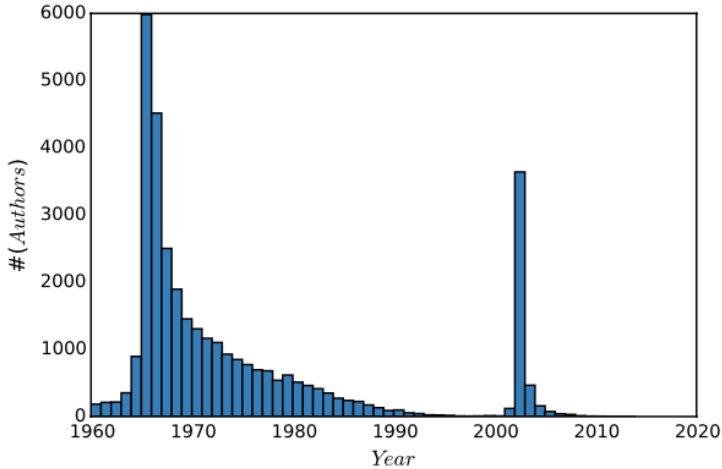
- Xác định vấn đề lỗi có quy tắc nào không -> chỉnh sửa lại theo 1 format nhất định

# Lỗi và giả tượng

- **Lỗi dữ liệu** đến từ mất mát trong quá trình thu thập,
  - Không thể khôi phục lỗi dữ liệu.
- Còn **giả tượng** đến từ sai sót trong quá trình xử lý dữ liệu
  - Có thể sửa chữa giả tượng miễn phát hiện và còn dữ liệu thô ban đầu
- Phát hiện giả tượng:
  - Xây dựng một mô hình kỳ vọng
  - Sniff test: Dò tìm “bad smell” – các điểm bất thường so với mô hình
  - Đặc biệt chú ý các tin tốt đáng ngờ

# Lỗi và giả tượng

Đếm số lượng tác giả khoa học xuất sắc theo năm lần đầu xuất hiện




Tìm các yếu tố bất thường có khả năng là giả tượng trong hình trên

# Lỗi và giả tượng

**Giả định:**

- Cùng với dân số, số tác giả có thể tăng nhẹ ở một đoạn nào đó, rồi giảm dần ở các năm gần trước 2020 (vì chưa tích đủ thành tác giả xuất sắc)
- Do mỗi lớp sinh viên tốt nghiệp đều thêm vào một số nhà khoa học, đồ thị sẽ phải khá phẳng.

**Quan sát:**

- Có 2 điểm chóp, một ở năm 1965, một ở 2002
- Trước đó đều tăng dốc, sau đó đều giảm dốc

**Giải thích:**

- Chóp trái: năm 1965, Pubmed bắt đầu thu thập các xuất bản khoa học
- 1960-1964: một số ít thu thập lẻ tẻ chưa hệ thống
- Không thể giải thích điểm chóp phải, độ dốc cao, sự giảm đến gần 0 rồi lại tăng => nhiều khả năng đều là giả tượng

# Lỗi và giả tượng

**Giải thích giả tượng:**

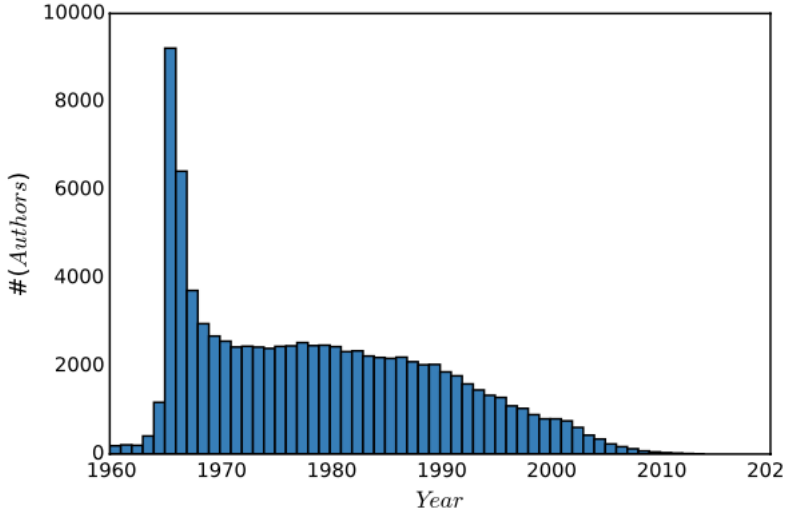
- Cuối 2001, Pubmed thay đổi cách viết tên tác giả.
- Một số tác giả đã tính ở năm trước đến 2022 có tên mới được tính lại
- Một số bị cắt mất phần nghiên cứu trước 2021 và không lọt top xuất sắc
- Rất ít số chỉ tính theo tên mới từ 2001 mà lọt top xuất sắc

**Sửa chữa giả tượng:**

- Thống nhất hệ thống tên gọi
- Thống kê lại 100000 tác giả xuất sắc nhất
- Vẽ lại đồ thị đếm lần đầu xuất hiện của 100000 tác giả này theo thời gian

# Lỗi và giả tượng

Sửa lại đồ thị số lượng tác giả khoa học xuất sắc theo năm lần đầu xuất hiện



# Sự không tương thích

"Không ai so táo với cam" – So các giá trị không cùng loại là vô nghĩa

- So 2 trọng lượng cùng giá trị mà khác đơn vị
- So doanh thu 2 bộ phim ở 2 thời điểm có giá trị đồng đô khác nhau: Endgame, Lord of the Rings, Avatar, Titanic, Gone with the Wind, …
- So giá vàng vào buổi trưa ở London và New York, khác múi giờ
- So giá cổ phiếu Microsoft ngày 17/2/03 với ngày 18/2/03 khi đợt chia tách làm giá giảm một nửa song tổng giá trị giữ nguyên

Cần chuyển các giá trị về cùng một nền tảng để so sánh

# Sự không tương thích

- **Chuyển đơn vị:**
  - Đơn vị chung nên là dạng chuẩn quốc tế (m, l, kg)
  - Cần đảm bảo biết mỗi trường nên dùng đơn vị gì để kiểm tra lại
  - Chú ý các bất thường trong dữ liệu, có thể chỉ do bị để ở đơn vị khác
  - Nếu trộn vài nguồn dữ liệu, tạo trường nguồn gốc để truy vết về sau
- **Chuyển biểu diễn số**
  - Hiểu về định dạng biểu diễn số nguyên, số thực, số bit biểu diễn
  - Không nên làm tròn số thực, sẽ sinh ra giả tượng

# Sự không tương thích

- **Thống nhất tên**
  - Chọn 1 cách biểu diễn tên và một cách mã hoá ký tự thống nhất
  - Hợp nhất quá quyết liệt cũng có rủi ro nhiều người thành 1 tên, nên ở mức nào là phải tuỳ bài toán
- **Thống nhất ngày giờ**
  - Khác biệt giữa các thời điểm, giữa các góc nhìn, các vùng địa lý
- **Thống nhất tiền tệ**
  - Khác biệt giữa tỷ suất quy đổi, do tỷ lệ lạm phát

# Dữ liệu bất thường

Là giá trị ngoại lai (outlier): quá khác so với phần còn lại. VD: 300 tuổi.

- Nhiều khả năng sinh ra do lỗi khi nhập liệu hay thu thập dữ liệu
- Tuy ít vẫn làm lệch các thông số thống kê, đặc biệt khi dữ liệu hiếm.
- Kiểm tra xem giá trị min, max có quá cách xa các giá trị khác không, tốt nhất bằng vẽ biểu đồ tần suất và khảo sát vị trí các điểm cực trị
- Xử lý bằng cách xoá bỏ, song tốt hơn nữa là có thể tìm ra vấn đề có tính hệ thống mà điểm bất thường mới là 1 chỉ báo ban đầu.

# Bài tập

| STT | Số tiền |
|---|---|
| 1 | 12 |
| 2 | 13 |
| 3 | 14 |
| 4 | 21 |
| 5 | 17 |
| 6 | 20 |
| 7 | 99 |
| 8 | 01 |
| 9 | 15 |
| 10 | 23 |

- Đâu là outlier(s)?

# Outline

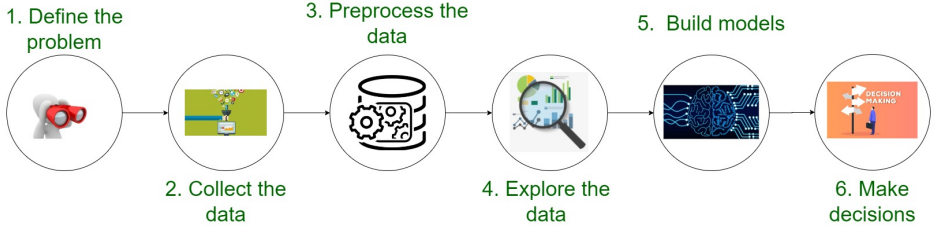
Tiền xử lý dữ liệu (tiếp)

**Phân tích cơ bản**

Thống kê mô tả

# Quy trình khoa học dữ liệu

Phân tích
Thống kê
Trực quan hoá
=> tổng quan



# Exploratory Data Analysis

Khảo sát đánh giá để tìm ra các khuôn mẫu, tính chất của dữ liệu có khả năng giải quyết vấn đề

- **Kỹ thuật phân tích cơ bản:** tính điểm, xếp hạng, min, max, sắp xếp, gộp nhóm
- **Kỹ thuật thống kê:** trung bình, trung vị, tứ phân vị, phương sai, độ lệch chuẩn, độ lệch, …

# Tính điểm (scoring)

Tính điểm giúp quy giản một thông tin nhiều chiều về một giá trị, làm nổi lên khía cạnh nào đó của thông tin.

VD 1: Từ các điểm bài về nhà, kiểm tra, thi quá trình, thi cuối kỳ, tính ra điểm cuối kỳ, 1 con số đại diện cho khả năng học môn đó của sinh viên.

- Đầy tính ngẫu nhiên: Mỗi môn, mỗi giáo viên một công thức
- Thiếu cách xác thực: Không ai chắc đâu là công thức "chính xác"
- Song nói chung vững chắc: Đa phần các công thức khác nhau cho ra xếp hạng như nhau.

Khó xác thực hoàn toàn hiệu quả, song vẫn là một phương pháp đánh giá dữ liệu hữu ích, đơn giản và phản ánh kinh nghiệm

# Ví dụ: Chỉ số BMI (Body Mass Index)

BMI = khối lượng (kg) / bình phương chiều cao (cm)

Là một chỉ số dùng để xếp hạng:

- Dưới 18.5: Gầy
- Từ 18.5 đến 25: Bình thường
- Từ 25 đến 30: Béo
- Trên 30: Béo phì

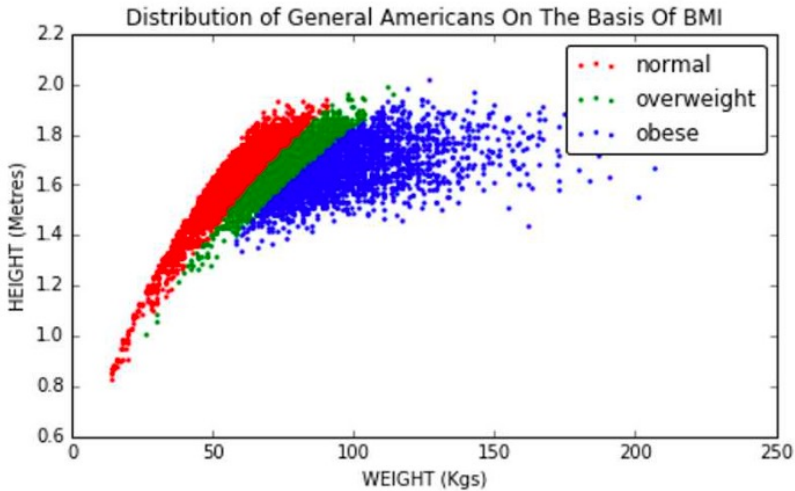
Được chấp nhận rộng rãi như một độ đo chuẩn để đánh giá gầy, béo

**Câu hỏi:**

Sao không dùng tỷ lệ mỡ trong cơ thể - độ đo chính xác về gầy, béo?

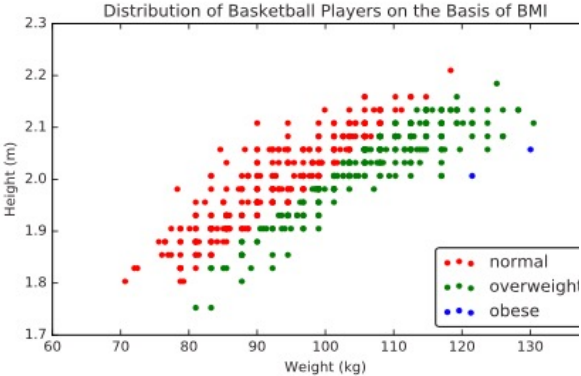
Tại sao lại bình phương mà không phải lập phương (tỷ lệ với thể tích)

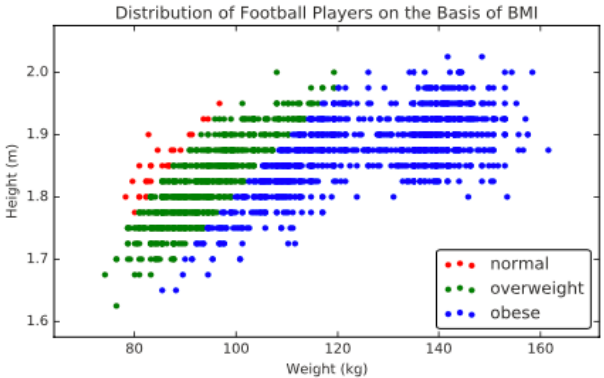
# BMI của người Mỹ




Các điểm trù mật và các điểm thưa thớt phản ánh những thông tin nào?

# BMI của vận động viên bóng rổ và bóng đá Mỹ




Nhận xét gì về hình trái, hình phải, hình trái vs. hình phải?

# Xây dựng một hệ thống tính điểm

- "Gold standard" (tiêu chuẩn vàng): đáp án cho vấn đề ta quan tâm.
- Nếu tồn tại tiêu chuẩn vàng thì có các cách hiệu quả để tính điểm. Coi điểm như là dự đoán giá trị của tiêu chuẩn vàng, rồi dùng các mô hình dự đoán như hồi quy tuyến tính để tìm công thức tính điểm.
- Không dễ tìm ra tiêu chuẩn vàng, thường dùng proxy (tiêu chuẩn đại diện) là các dữ liệu dễ tìm hơn, tương quan mạnh với tiêu chuẩn vàng. VD: Dùng BMI làm proxy cho tỷ lệ mỡ.
- **Câu hỏi:** Dùng gì làm proxy để đánh giá cách tính điểm một môn học?

# Đánh giá một hệ thống tính điểm

- **Dễ tính toán:** BMI chỉ 2 biến, số học đơn giản, các biến đều dễ đo đạc.
- **Dễ hiểu:** Có sự liên hệ tự nhiên giữa chiều cao, cân nặng, và gầy, béo
- **Đơn điệu theo biến:** đồng biến với cân nặng, nghịch biến với chiều cao.
- **Kết quả tốt cho giá trị ngoại lai:** BMI lớn với thừa cân, nhỏ với thiếu cân
- **Chuẩn hoá từng biến:** hay dùng chuẩn hoá kiểu z-score

$$Z_i = (a_i - \mu)/\sigma$$

- **Chọn tiêu chí phân loại ý nghĩa:** làm sao để phân loại được

**Bài tập:** Xây dựng công thức tính điểm quá trình cho môn Nhập môn DS

Giả định có 5 điểm bài tập, 2 điểm quá trình, sinh viên sẽ được biết điểm cả lớp vào thứ 5 tuần 9. Deadline nộp công thức tính điểm, thứ 2 tuần 10.

# Xếp hạng (Ranking)

Là một phép hoán vị *n* thực thể theo một tiêu chí sắp xếp nào đó. Xếp hạng được xây dựng trên một hệ thống tính điểm.

- Xếp hạng top đội bóng/cầu thủ
- Xếp hạng học thuật các trường đại học
- Xếp hạng kết quả tìm kiếm
- Xếp hạng học sinh
- Xếp hạng bài hát

Xếp hạng cũng gắn một thực thể với một con số (hạng). Giữa 2 con số, hạng và điểm, đâu là độ đo ý nghĩa hơn trong đánh giá dữ liệu?

# Xếp hạng vs. Tính điểm

- Thông tin có thể độc lập hay cần có ngữ cảnh?
- Hệ thống tính điểm có phổ cập và dễ hiểu?
- Có quan tâm tới khoảng cách giữa các thực thể?
- Có quan tâm tới phân phối tổng thể của mọi thực thể?
- Quan tâm tới các điểm cực trị hay các điểm ở trung tâm?

# Outline

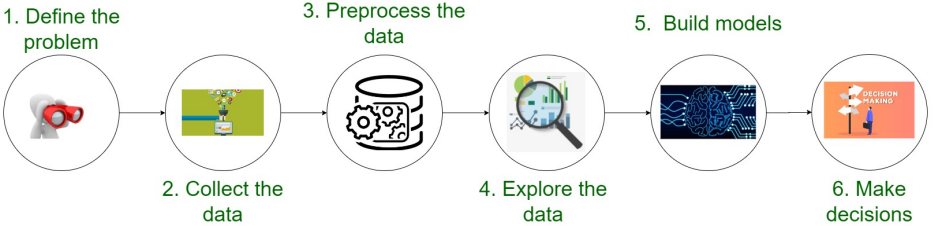
Tiền xử lý dữ liệu (tiếp)

Phân tích cơ bản

**Thống kê mô tả**

# Quy trình khoa học dữ liệu

Phân tích
Thống kê
Trực quan hoá
=> tổng quan



# Exploratory Data Analysis

Khảo sát đánh giá để tìm ra các khuôn mẫu, tính chất của dữ liệu có khả năng giải quyết vấn đề

- **Kỹ thuật phân tích cơ bản:** tính điểm, xếp hạng, min, max, sắp xếp, gộp nhóm
- **Kỹ thuật thống kê:** trung bình, trung vị, tứ phân vị, phương sai, độ lệch chuẩn, độ lệch, …

# Thống kê mô tả

- Giúp quy giản một lượng lớn dữ liệu theo một cách hợp lý. Không nhằm dự đoán hay kiểm định giả thuyết như thống kê suy diễn, chỉ để mô tả.
- Trình bày dữ liệu dưới dạng các giá trị định lượng dễ quản lý:

Dựa trên 2 khái niệm chính

- Quần thể (population): tập hợp đối tượng mà ta tìm kiếm thông tin
- Mẫu (sample): tập con của quần thể mà ta quan sát được.

Ta áp dụng các khái niệm, độ đo, thuật ngữ thống kê thu được các đặc trưng tổng quan về mẫu, và dùng chúng để ước lượng quần thể.

# Các khái niệm thống kê

- Tần suất
- Phân phối, phân phối tích luỹ, họ phân phối chuẩn
- Kỳ vọng/trung bình
- Phương sai, độ lệch chuẩn
- Trung vị, tứ phân vị và bách phân vị
- Giá trị ngoại lai
- Độ lệch và độ lệch Pearson
- Ước lượng tham số

# Khối lượng xác suất, mật độ xác suất, và phân phối tích luỹ

- **Tần suất/xác suất:** số lần xuất hiện giá trị đó trên toàn quần thể. Có thể chuẩn hoá thành tỷ lệ xuất hiện giá trị trên toàn quần thể.
- **Hàm khối lượng xác suất** (Probability Mass Function - PMF): đo tần xuất/xác suất xuất hiện của một biến rời rạc.

$$P(\mathrm{x} = x) \geq 0 \ \forall x \in \mathrm{X} \text{ và } \sum_{x \in \mathrm{X}} P(\mathrm{x} = x) = 1$$

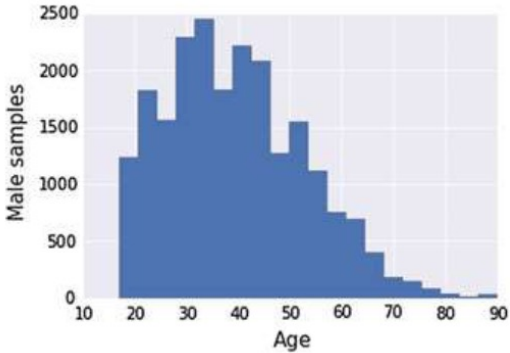
- **Hàm mật độ xác suất** (Probability Mass Function - PDF) đo mật độ của một biến liên tục, lấy tích phân khoảng cho xác suất biến rơi vào khoảng đó:

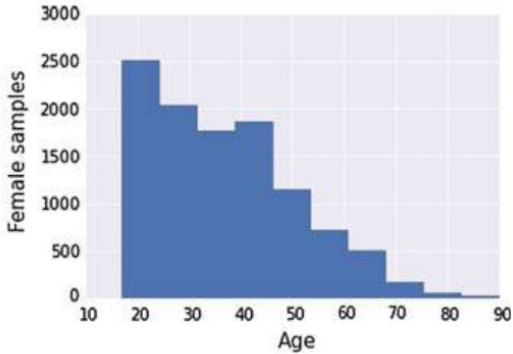
$$p(\mathrm{x} = x) \geq 0 \ \forall x \in \mathrm{X} \text{ và } \int_{\mathrm{X}}^{\infty} p(\mathrm{x} = x)dx = 1$$

- **Hàm phân phối (tích luỹ)** (Cumulative Distribution Function - CDF) đo xác suất xuất hiện một giá trị nhỏ hơn hay bằng giá trị *x* trong quần thể, bằng tổng các PMF hoặc tích phân của PDF.

# Ví dụ hàm khối lượng xác suất

Phân phối tuổi lao động của đàn ông và phụ nữ

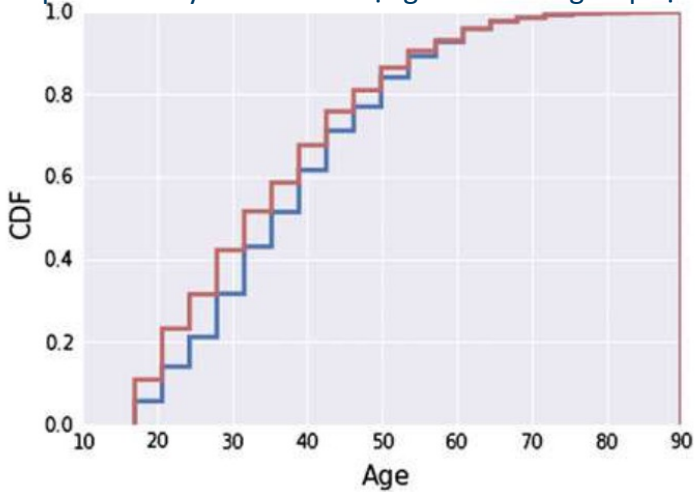



Nhận xét gì?

# Ví dụ hàm phân phối tích luỹ

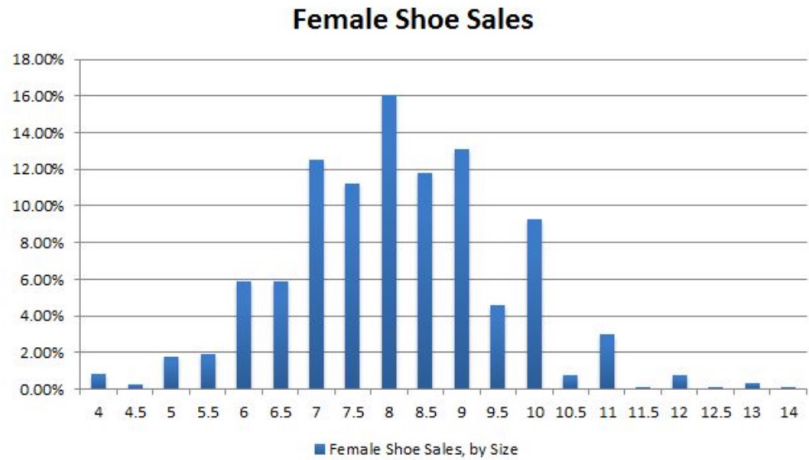
Phân phối tích luỹ về tuổi lao động của đàn ông và phụ nữ




Nhận xét gì?

# Ví dụ hàm mật độ xác suất

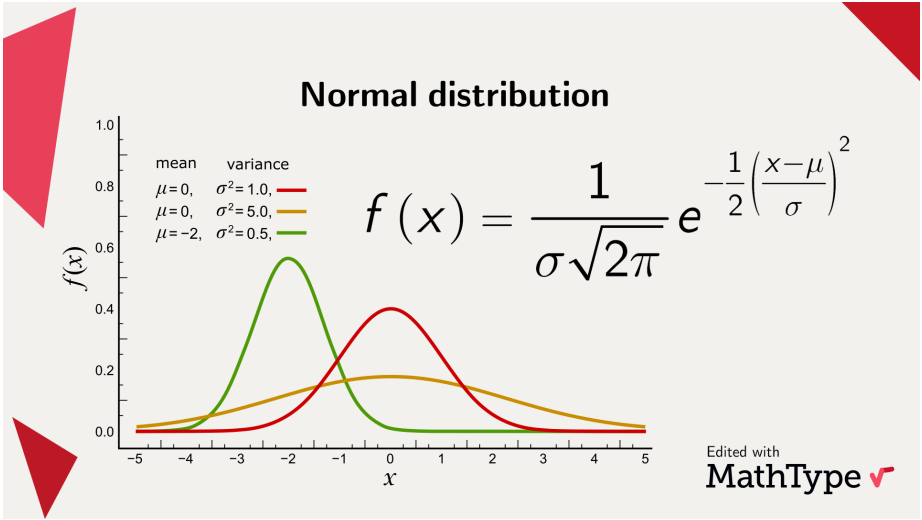
Phân phối doanh thu theo cỡ giày của phu nữ




Họ hàm mật độ xác suất phổ thông nhất là họ phân phối chuẩn - phân phối Gausse – phân phối hình chuông

# Họ phân phối cơ bản: Phân phối chuẩn/Gausse

- Dùng cho biến có giá trị liên tục



# Kỳ vọng, phương sai, độ lệch chuẩn

- **Kỳ vọng:** thể hiện giá trị trung tâm dữ liệu, bằng trung bình các giá trị

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- **Phương sai:** thể hiện độ phân tán của dữ liệu, tính bằng trung bình bình phương khoảng cách tới kỳ vọng,

$$\sigma^2 = \frac{1}{n}\sum_{i}(x_i - \mu)^2.$$

- **Độ lệch chuẩn:** Căn của phương sai

# Trung vị, tứ phân vị, bách phân vị

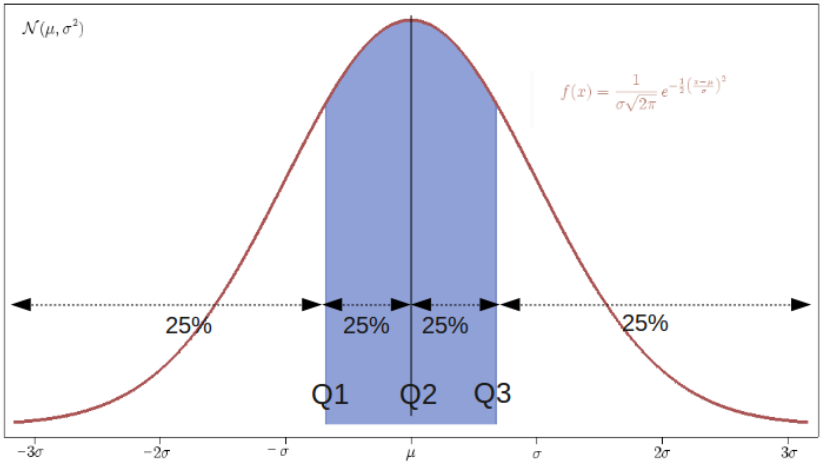
Có thể chia quần thể thành 2 tập con dựa theo một giá trị xác suất p, với điểm chia ký hiệu là $x_p$ như sau:

- Số giá trị nhỏ hơn hay bằng $x_p$ sẽ chiếm tỷ lệ *p* trong mẫu
- Số giá trị lớn hơn $x_p$ chiếm tỷ lệ *1-p* trong mẫu

Từ đây suy ra các khái niệm:

- **Trung vị:** điểm chia đôi phân phối xác suất, ký hiệu $\mu_{12}$. Trung vị ổn định hơn và không bị kéo lệch bởi giá trị ngoại lai như kỳ vọng.
- **Tứ phân vị:** các điểm chia tư xác suất, ký hiệu *Q1, Q2, Q3.*
- **Bách phân vị:** các điểm chia 100 xác suất (nhóm xếp hạng 10, 30, 50)

# Ví dụ: Trung vị, tứ phân vị, bách phân vị




Điểm trung vị là điểm tứ phân vị nào? Các điểm trung vị, tứ phân vị là các điểm bách phân vị nào?

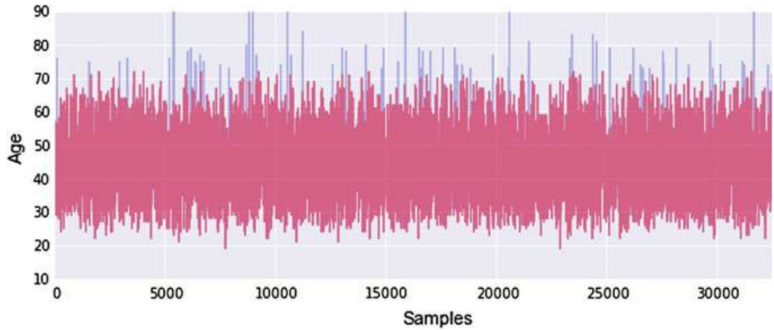
# Giá trị ngoại lai/bất thường

Các giá trị quá khác biệt so với phần còn lại của dữ liệu. Trên quan điểm thống kê ta có thể kiểm tra giá trị ngoại lai như sau:

- Các mẫu nằm quá xa trung vị
- Các mẫu lệch so với kỳ vọng hơn 2 lần (hay 3 lần) độ lệch chuẩn

Loại bỏ giá trị ngoại lai sẽ kéo giá trị trung bình về gần hơn trung vị.

VD: màu tím nhạt là các giá trị ngoại lai



# Bài tập

| STT | Số tiền |
|---|---|
| 1 | 12 |
| 2 | 13 |
| 3 | 14 |
| 4 | 21 |
| 5 | 17 |
| 6 | 20 |
| 7 | 99 |
| 8 | 01 |
| 9 | 15 |
| 10 | 23 |

- Tìm kỳ vọng, phương sai, độ lệch chuẩn, điểm trung vị
- Tìm dữ liệu outlier trong hình này bằng phương pháp thống kê (2 cách):
  + cách 1: lệch so với kỳ vọng quá 2 độ lệch chuẩn

**Về nhà:**

+ cách 2: lệch so với trung vị quá 50%

- Tính lại các giá trị kỳ vọng, phương sai, độ lệch chuẩn, điểm trung vị, tứ phân vị cho cả 2 trường hợp bỏ outlier:
- Nhận xét kết quả

# Độ lệch

Các giá trị quá khác biệt so với phần còn lại của dữ liệu. Trên quan điểm thống kê ta có thể kiểm tra giá trị ngoại lai như sau:

- Các mẫu nằm quá xa trung vị
- Các mẫu lệch so với kỳ vọng hơn 2 lần (hay 3 lần) độ lệch chuẩn

Loại bỏ giá trị ngoại lai sẽ kéo giá trị trung bình về gần hơn trung vị.

Công thức và ký hiệu độ lệch:

$$\tilde{\mu}_3 = \mathrm{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right]$$

# Hệ số lệch Pearson

**Formula for Pearson's Skewness**

$$Sk_1 = \frac{\bar{X} - Mo}{s}$$

---

$$Sk_2 = \frac{3\bar{X} - Md}{s}$$

**where:**

$Sk_1 =$ Pearson's first coefficient of skewness and $Sk_2$ the second

$s =$ the standard deviation for the sample

$\bar{X} =$ is the mean value

$Mo =$ the modal (mode) value

$Md =$ is the median value

Sk2 không dùngmode, nên có thể dùng với mô hình có mode yếu hay nhiều mode

# Hiệp phương sai và hệ số tương quan Pearson

- Đo mức độ tương quan hay quan hệ giữa 2 đại lượng, liệu có thể dùng đại lượng này để dự đoán đại lượng kia
- Hiệp phương sai (covariance)

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} dx_i dy_i,$$

- Hệ số tương quan Pearson:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

# Ước lượng tham số

- Về cơ bản ta ước lượng các tham số của quần thể (kỳ vọng, phương sai, …) bằng công thức tương ứng song tính trên tập mẫu. VD coi kỳ vọng quần thể tính từ kỳ vọng mẫu, nghĩa là bằng trung bình các giá trị trong tập mẫu.
- Song khi tập mẫu quá bé, vài tham số mẫu cần thay đổi công thức:
  - Phương sai mẫu với dữ liệu bé:

$$\bar{\sigma}^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2.$$

  - Độ lệch mẫu với dữ liệu bé:

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

# Tổng kết

- Các vấn đề tiềm ẩn: thiếu hụt (thiếu giá trị, trùng lặp, sai logic), sai sót (lỗi mất mát, giả tượng), không tương thích (đơn vị, ngày giờ, tài chính, biểu diễn số, biểu diễn tên), giá trị ngoại lai/khác thường
- Các cách làm sạch dữ liệu để xử lý vấn đề tiềm ẩn
- Quy trình EDA: phân tích cơ bản (chấm điểm và xếp hạng) và thống kê mô tả